Capturing and Interpreting Unique Information

Praveen Venkatesh
Allen Institute
and University of Washington
Seattle, WA, USA
praveen.venkatesh@alleninstitute.org

Keerthana Gurushankar
Department of Computer Science,
Carnegie Mellon University
Pittsburgh, PA, USA
kgurusha@andrew.cmu.edu

Gabriel Schamberg
Department of Surgery,
University of Auckland
New Zealand
gabe.schamberg@auckland.ac.nz

***Abstract*—Partial information decompositions (PIDs), which quantify information interactions between three or more variables in terms of uniqueness, redundancy and synergy, are gaining traction in many application domains. However, our understanding of the operational interpretations of PIDs is still incomplete for many popular PID definitions. In this paper, we discuss the operational interpretations of unique information through the lens of two well-known PID definitions. We reexamine an interpretation from statistical decision theory showing how unique information upper bounds the risk in a decision problem. We then explore a new connection between the two PIDs, which allows us to develop an informal but appealing interpretation, and generalize the PID definitions using a common Lagrangian formulation. Finally, we provide a new PID definition that is able to *capture* the information that is unique. We also show that it has a straightforward interpretation and examine its properties.**

The full version of this paper is available online [1].

I. Introduction

Partial information decompositions (PIDs) have become a popular method for understanding the information interactions between multiple random variables. A *bivariate* PID seeks to decompose the information that two variables X and Y convey about a message M, into parts that are unique to X, unique to Y, redundant to X and Y, and synergistic [2]–[4].

As a simple example, consider a message $M = [M_1, M_2, M_3, M_4]$, and two variables $X = [M_1, M_3, M_4 \oplus Z]$ and $Y = [M_2, M_3, Z]$, where $M_i, Z \sim$ i.i.d. Ber$(1/2)$ and $\oplus$ represents an XOR operation between bits. Here, X has one bit of unique information about M, i.e., M_1, which is not present in Y. Similarly, Y has one bit of unique information about M, i.e., M_2, which is not present in X. There is one bit of redundant information, i.e., M_3, which can be extracted from *either* X or Y taken alone. Finally, there is one bit of synergistic information, i.e., M_4: this information cannot be extracted from either X or Y individually, but can be recovered when both are taken *together*.

PIDs have found applications in various fields, from neuroscience [5], [6] (where one may want to examine the interaction between stimuli, neural activity and behavioral response) to financial markets [7]. PIDs have also been used to explain how information complexity decreases through the layers of a deep neural network [8], as well as by us to develop new measures of fairness in machine learning [9].

Despite increasingly widespread adoption, there is still no consensus on how PIDs should be defined, and even more so, on how to operationally interpret partial information quantities (e.g., see [10], [11]). In this paper, we focus on interpretations of unique information. What do we mean by an operational interpretation here? In essence, we would like to make formal a statement of the variety "X has access to u bits of information about M that Y does not have access to". A concrete operational interpretation would mathematically define terms like "having access" within a certain context.

One popular approach for operational interpretations has relied on the concept of Blackwell sufficiency from statistical decision theory. Blackwell sufficiency is a formal way to determine whether X contains all of the information that Y has about M. Thus, it becomes a natural basis for discussing how two variables carry information about a message.[1] Bertschinger et al. [4] used Blackwell sufficiency to motivate a definition of unique information. But their interpretation only addressed whether the unique information was zero or non-zero, and did not provide an interpretation for the *quantity* of unique information. More recently, Banerjee et al. [3] and Rauh et al. [12] interpreted the quantification of unique information as an upper bound on a "secret key rate", which is well-defined in the context of information-theoretic security. However, such an interpretation is less well-defined other contexts like neuroscience, where there is no clear analog for an eavesdropper or a secret key.

In this paper, we consider two PID definitions based on Blackwell sufficiency [3], [4], and discuss a more broadly applicable operational interpretation of the *quantity* of unique information in each case. Extending classical results on so-called "deficiency" measures [13], [14], it can be shown that the unique information about M present in X w.r.t. Y upper bounds the difference in risk attained in a decision problem, when one uses X rather than Y to make decisions pertaining to M. This interpretation was also stated in passing by Banerjee et al. [3]; we believe this is a more broadly applicable interpretation, and hence place renewed emphasis on it here (Sections III-A and III-B). In the process, we also explicitly discuss gaps in our understanding that are yet to be filled (Section III-C).

We then identify a previously unrecognized connection between the aforementioned PIDs, which shows that the two definitions swap the objective and constraint in their respective optimizations (Section III-D). This discovery allows us to clarify how these definitions are related to Blackwell sufficiency, and provide an informal but appealing interpretation for them (Section III-E). Finally, we develop a novel generalization of

[1]For example, Kolchinsky [11] uses it to operationalize measures of redundancy and "union" information.

the two PIDs, through a common Lagrangian (Section III-F).

Going beyond operational interpretations, we would also like to know *what* the unique information *is*, not just how *much* there is. Towards this, in Section IV, we propose a new PID definition (which we had hinted at in our previous work [15]) that "captures" the *part* of M that is unique in the form of a random variable. We show that it forms a valid non-negative decomposition obeying intuitive bounds, and that it has a simple and appealing interpretation. We also show that this PID definition is Blackwellian [16] when M, X and Y are jointly Gaussian.

II. Background

A. Notation

- Let M, X and Y be three random variables with sample spaces M, X and Y respectively, and joint density P_{MXY}.
- Let $\mathcal{C}(\mathsf{A}\,|\,\mathsf{B})$ denote the set of all *channels* (conditional distributions) from A to B, so for example, $P_{X|M} \in \mathcal{C}(\mathsf{X}\,|\,\mathsf{M})$.
- Let $\circ$ denote composition of channels, i.e. $\forall\ a \in \mathsf{A}, c \in \mathsf{C}$,

$$(P_{A|B} \circ P_{B|C})(a\,|\,c) \coloneqq \int_{\mathsf{B}} P_{A|B}(a\,|\,b) \cdot P_{B|C}(b\,|\,c)\, db.$$

- To keep the exposition simple, we ignore any measure-theoretic nuances. All conditional distributions and information measures are assumed to be well-defined.

B. Defining PIDs

There are many notions of partial information decompositions: we focus here on the *bivariate* case, which decomposes the information that *two* variables X and Y have about a message M. Such a PID is typically defined by a set of four functions of the joint distribution P_{MXY}—denoted $UI(M : X \setminus Y)$, $UI(M : Y \setminus X)$, $RI(M : X; Y)$ and $SI(M : X; Y)$ (or UI_X, UI_Y, RI and SI respectively for brevity)—which satisfy the following basic equations:

$$\begin{aligned} I\big(M; (X,Y)\big) &= UI(M : X \setminus Y) + UI(M : Y \setminus X) \\ &\quad + RI(M : X; Y) + SI(M : X; Y), \quad (1) \\ I\big(M; X\big) &= UI(M : X \setminus Y) + RI(M : X; Y), \quad (2) \\ I\big(M; Y\big) &= UI(M : Y \setminus X) + RI(M : X; Y). \quad (3) \end{aligned}$$

Equation (1) implies that the total mutual information about M conveyed by X and Y is the sum of four partial information components: one unique to X, one unique to Y, another redundant to both X and Y, and the last which is synergistic, respectively. Equations (2) and (3) enforce that the individual mutual information of X or Y with M is the sum of the redundant information and the corresponding unique information.[2] These equations impose three constraints on the four partial information components, such that defining any one component suffices to specify the other three.

In this paper we discuss the operational interpretations of two existing PID definitions due to [3] and [4] in Section III, and then introduce a new PID definition in Section IV. We begin by stating the first two definitions, and defining the concept of Blackwell sufficiency upon which they are based.

Definition 1 (δ-PID [3])**.** *Let the (weighted output)* deficiency[3] *of Y with respect to X about M be defined as*[4]

$$\delta(M : X \backslash Y) \coloneqq \inf_{P_{X'|Y} \in \mathcal{C}(\mathsf{X}|\mathsf{Y})} \mathbb{E}_{P_M}\big[D(P_{X|M} \,\|\, P_{X'|Y} \circ P_{Y|M})\big]. \quad (4)$$

Then, the deficiency-based redundant information about M present in X and Y is given by

$$\begin{aligned} RI^{\delta}(M : X; Y) \coloneqq \min\{&I(M;X) - \delta(M : X \setminus Y), \\ &I(M;Y) - \delta(M : Y \setminus X)\}. \end{aligned} \quad (5)$$

Using equations (1)–(3), RI^{δ}_X fully determines the δ-PID, i.e. UI^{δ}_X, UI^{δ}_Y, and SI^{δ}.

Definition 2 ($\sim$-PID[5] [4], [17])**.** *The unique information about M present in X and not in Y is given by*

$$\widetilde{UI}(M : X \setminus Y) \coloneqq \min_{Q \in \Delta_P} I_Q(M; X\,|\,Y), \quad (6)$$

where $\Delta_P \coloneqq \{Q_{MXY} : Q_{MX} = P_{MX},\ Q_{MY} = P_{MY}\}$ and $I_Q(\cdot\,|\,\cdot)$ is the conditional mutual information over the joint distribution Q_{MXY}.

As with the δ-PID, equations (1)–(3) fully determine the remaining components of the $\sim$-PID.

C. Blackwell sufficiency and Blackwellian PIDs

Blackwell sufficiency provides a partial order between random variables based on how informative they are about a message M. This notion was used by [4] to provide an operational motivation for the $\sim$-PID, and also underlies the basis of the δ-PID [3].

Definition 3 (Blackwell sufficiency: $\succcurlyeq_M$)**.** *We say that a channel $P_{X|M}$ is* Blackwell sufficient *w.r.t. another channel $P_{Y|M}$ (denoted $X \succcurlyeq_M Y$) if $\exists\ P_{Y'|X} \in \mathcal{C}(\mathsf{Y}\,|\,\mathsf{X})$ such that*

$$P_{Y'|X} \circ P_{X|M} \;=\; P_{Y|M}. \quad (7)$$

Intuitively, $X \succcurlyeq_M Y$ means that we can generate a new random variable Y' from X (using the stochastic transformation $P_{Y'|X}$) so that the effective channel from M to Y' is equivalent to the original channel from M to Y.[6] It was shown by Blackwell [18] that if X is Blackwell sufficient for M w.r.t. Y, then it is always preferable to observe X rather than Y, for making decisions about M. This operational interpretation of Blackwell sufficiency was extended to PIDs by [4]:

Definition 4 (Blackwellian PID)**.** *A bivariate PID on P_{MXY} is said to be Blackwellian if*

$$UI_X = 0 \;\Leftrightarrow\; Y \succcurlyeq_M X \quad \textit{and} \quad UI_Y = 0 \;\Leftrightarrow\; X \succcurlyeq_M Y$$

[2] Typically, it is also assumed that the redundant and synergistic components are symmetric in X and Y.

[3] *Deficiency* was introduced by Le Cam to quantify a departure from Blackwell *sufficiency*.

[4] The reason for this notation is that the *deficiency* of Y w.r.t. X translates to the *unique information* present in X and not in Y.

[5] Also called the BROJA-PID in the literature after the authors of [4].

[6] Blackwell sufficiency is identical to the concept of stochastic degradedness of broadcast channels [16].

This means that (for a Blackwellian PID definition) the unique information in one variable is zero only if it is always beneficial to observe the *other* variable to make decisions about M. Conversely, if X is not Blackwell sufficient for M w.r.t. Y, then Y must have some unique information about M that X cannot access.

However, it is important to note that a Blackwellian PID is only operationally motivated to the extent of whether or not the unique information is *zero*. It does not lend an operational interpretation as to the *volume* of unique information when it is non-zero.

III. Interpreting the δ- and $\sim$-PIDs

A. Deficiency upper bounds the difference in risk

The δ-PID derives its operational interpretation directly from that of deficiency [13], [19], upon which it is based. The deficiency of Y w.r.t. X, originally defined by Le Cam [19], measures how far from Blackwell sufficient Y is, w.r.t. X.

Le Cam's original notion of deficiency was defined using the total variation distance, and as a worst case over realizations of M. That was a frequentist context, where M was a statistical parameter and not a random variable. Following Raginsky [20], the Le Cam deficiency of Y w.r.t. X about M is:

$$\begin{aligned}&\delta^{\text{LeCam}}(M : X \setminus Y)\\ &\quad\coloneqq \inf_{\substack{P_{X'|Y}\\ \in \mathcal{C}(X|Y)}} \sup_{m\in\mathsf{M}} \left\|P_{X'|Y}\circ P_{Y|M=m} - P_{X|M=m}\right\|_{TV}\end{aligned} \tag{8}$$

The Le Cam deficiency can be interpreted as upper bounding the difference in risk (for any bounded loss function) when using X rather than Y to make decisions based on M. We can state this formally, using the setup of a decision problem:

Definition 5 (Decision problem)**.** *Suppose we need to perform actions based on the value of M, which we cannot observe directly (e.g., we may want to estimate the value of M). We have access to either $X \sim P_{X|M}$ or $Y \sim P_{Y|M}$, which can give us information about M. The actions we take after observing either X or Y—call these $\widehat{M}_X(x)$ and $\widehat{M}_Y(y)$ respectively—incur a bounded loss that depends on the chosen action and the value of M. Let $\mathcal{L}(\widehat{M}(\cdot), M)$ ($\|\mathcal{L}\|_\infty \leq 1$) be the loss function, where $\widehat{M}(\cdot)$ may be either $\widehat{M}_X(x)$ or $\widehat{M}_Y(y)$, depending on whether we choose to observe X or Y. How do we decide whether to choose X or Y when we do not know $\mathcal{L}$?*

Blackwell [18] showed that if $X \succcurlyeq_M Y$, we can always attain a lower expected loss by choosing X. What happens when Blackwell sufficiency does not hold? Define the risk as the expected loss over either X or Y:

$$\mathcal{R}_m(P_{X|M}, \widehat{M}_X, \mathcal{L}) \coloneqq \mathbb{E}_{X\sim P_{X|M=m}}\left[\mathcal{L}(\widehat{M}_X(X), m)\right] \tag{9}$$

If Blackwell sufficiency does not hold, then the worst-case risk (over M) when you choose X is at most that when you choose Y, plus the Le Cam deficiency of X [13], [14]. In other words, for any m and for any $\widehat{M}_Y$, there exists an $\widehat{M}_X$ such that[7]

$$\begin{aligned}\mathcal{R}_m(P_{X|M}, \widehat{M}_X, \mathcal{L}) \leq \mathcal{R}_m(P_{Y|M}, \widehat{M}_Y, \mathcal{L})\\ + \delta^{\text{LeCam}}(M : Y \setminus X).\end{aligned} \tag{10}$$

Raginsky [20] showed how alternative measures like the KL-divergence may be used in place of the total variation distance in (8), while preserving the aforementioned risk-based operational interpretation. In that work, Raginsky preserved the frequentist setting, taking the worst case divergence between $P_{X'|Y}\circ P_{Y|M=m}$ and $P_{X|M=m}$, over all realizations of M. However, for PIDs, M is a random variable and thus it makes more sense to consider the *expected* divergence over different values of M. This is what Banerjee et al. [3] did, in proposing the weighted output deficiency stated in Definition 1. They also showed that the decision-theoretic operational interpretation extends to the new deficiency definition $\delta(M : X \setminus Y)$ [3, Prop. 8]. We restate this theorem here, and provide a proof in Appendix A [1] for completeness.

Theorem 1 (Prop. 8 in [3])**.** *Let the average risk be given by*

$$\bar{\mathcal{R}}(P_{X|M}, \widehat{M}_X, \mathcal{L}) \coloneqq \mathbb{E}_{M,X}\left[\mathcal{L}(\widehat{M}_X(X), M)\right] \tag{11}$$

Then, for any $\widehat{M}_Y$, there exists an $\widehat{M}_X$ such that

$$\begin{aligned}\bar{\mathcal{R}}(P_{X|M}, \widehat{M}_X, \mathcal{L}) \leq \bar{\mathcal{R}}(P_{Y|M}, \widehat{M}_Y, \mathcal{L})\\ + g(\delta(M : Y \setminus X)),\end{aligned} \tag{12}$$

where $g(\cdot)$ is a monotonically increasing function.

B. UI^δ and $\widetilde{UI}$ upper bound the difference in risk

Despite the existence of a clear operational interpretation for deficiency as defined in Definition 1, the δ-PID, which arises out of deficiency, still needs an interpretation. In particular, we need to address what happens after we symmetrize the redundancy in Equation (5).[8]

Corollary 2. *UI^δ may be used in place of δ in Theorem 1.*

Proof. The unique information upper bounds the deficiency:

$$\begin{aligned}&UI^\delta(M : Y \setminus X)\\ &\quad= I(M;Y) - RI^\delta(M : X;Y) &(13)\\ &\quad= \max\{\delta(M : Y \setminus X),\\ &\qquad\qquad \delta(M : X \setminus Y) + I(M;Y) - I(M;X)\} &(14)\\ &\quad\geq \delta(M : Y \setminus X). &(15)\end{aligned}$$

The rest follows from the fact that $g(\cdot)$ is a monotonically increasing function. □

Thus, the decision-theoretic operational interpretation also applies to the unique information of the δ-PID, although the bound that it implies may be somewhat loose.

[7] Recall that the *deficiency in X* is denoted $\delta(M : Y \setminus X)$, because it corresponds to the *unique information in Y*.

[8] This symmetrization step is required because $I(M;X) - \delta(M : X \setminus Y)$ is not always symmetric in X and Y. Interestingly, this issue does not arise in the case of the $\sim$-PID, which has an intrinsically symmetric redundancy [4].

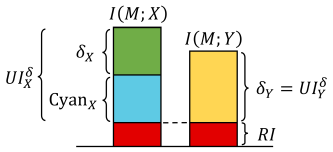

Fig. 1: A depiction of the cyan region problem described in Section III-C for the δ-PID. The two bars represent the quantity of mutual information M has with X and Y respectively; the green and yellow portions represent how much of that information is the deficiency; and the red portion represents the symmetrized redundancy. The cyan region is part of the unique information in X, but cannot be accounted for by deficiency.

The unique information of the $\sim$-PID, $\widetilde{UI}_Y$, also acts as an upper bound for the difference in risk when choosing X rather than Y in the decision problem from Definition 5.

Corollary 3. *$\widetilde{UI}$ may be used in place of δ in Theorem 1.*[9]

Proof. This follows directly from a result of Bertschinger et al. [4], which states that $\widetilde{UI}$ upper bounds the unique information of any other PID definition that satisfies what they call "Assumption $(*)$". According to this assumption, a definition of unique information should depend only on P_M, $P_{X|M}$ and $P_{Y|M}$, and not on the whole joint distribution P_{MXY}. Since the δ-PID satisfies Assumption $(*)$, we have that $\widetilde{UI}_Y \geq UI_Y^\delta$, which, along with Corollary 2, completes the proof. However, the upper bound may once again be loose. □

C. An unbridged gap in the decision-theoretic interpretation

For both UI^δ and $\widetilde{UI}$, the decision-theoretic operational interpretation does not yield a tight bound. In fact, one of two unique informations, UI_X or UI_Y, is guaranteed to be loose in this way. Taking the case of UI^δ, which has the tighter of the two bounds, we can quantify the extent of slack as follows: suppose that $I(M;X) - \delta(M : X \setminus Y) > I(M;Y) - \delta(M : Y \setminus X)$. Then, $RI^\delta(M : X;Y) = I(M;Y) - \delta(M : Y \setminus X)$, and thus

$$\begin{aligned} UI^\delta(M : Y \setminus X) &= \delta(M : Y \setminus X) \\ UI^\delta(M : X \setminus Y) &= \delta(M : Y \setminus X) + I(M;X) - I(M;Y). \end{aligned} \tag{16}$$

In other words, the excess quantity added to $UI^\delta(M : X \setminus Y)$, over and above the deficiency is

$$\begin{aligned} \text{Cyan}(M : X \setminus Y) \coloneqq I(M;X) - \delta(M : X \setminus Y) \\ - I(M;Y) + \delta(M : Y \setminus X). \end{aligned} \tag{17}$$

For lack of a better name, we call this the "cyan region", due to how it is depicted in Figure 1. It is completely unclear what the interpretation of $\text{Cyan}(M : X \setminus Y)$ ought to be, and why this information should be considered unique to X (see Figure 1).

Essentially, we pay the cost of a loose bound in $UI(M : X \setminus Y)$, and the extent of slack does not have a clear justification in and of itself. This is a gap in our decision-theoretic understanding the interpretation of unique information, which we leave to future work to fill.

[9]Corollaries 2 and 3 are implicit in [3]; we make these explicit here to emphasize the usefulness of this interpretation.

D. A connection between the $\sim$-PID and the δ-PID

We now present a previously unidentified connection between these two PIDs, and use this connection to develop an intuitive, albeit informal, interpretation for both PIDs.

First, observe that the δ-PID can be thought of as optimizing $P_{X'|MY}$ instead of $P_{X'|Y}$, so long as we include the constraint that M—Y—X' forms a Markov chain. This constraint can also be written as $I(M;X' \,|\, Y) = 0$. Thus, abbreviating $P_{X'|Y} \circ P_{Y|M}$ as $P_{X'|M}$, we can rewrite the deficiency from Equation (4) as:

$$\begin{aligned} \delta(M : X \setminus Y) = \inf_{P_{X'|MY}} \mathbb{E}_M \left[D_{KL}(P_{X|M} \,\|\, P_{X'|M}) \right] \\ \text{s.t.} \quad I(M;X' \,|\, Y) = 0. \end{aligned} \tag{18}$$

Next, we show that the definition of $\sim$-PID can also be rewritten into a similar form. The optimization variable Q in Definition 2 obeys the constraints that $Q_{MX} = P_{MX}$ and $Q_{MY} = P_{MY}$. Suppose we change notation by introducing a new random variable X' using the stochastic transformation $P_{X'|MY}$, but which also obeys $P_{X'M} = P_{XM}$—or equivalently, $P_{X'|M} = P_{X|M}$. Then, the distribution $P_{MX'Y}$ plays exactly the same role as Q_{MXY}, and obeys precisely the same constraints. Thus, the $\sim$-PID definition can also be written as:

$$\begin{aligned} \widetilde{UI}(M : X \setminus Y) = \inf_{P_{X'|MY}} I(M;X' \,|\, Y) \\ \text{s.t.} \quad \mathbb{E}_M \left[D_{KL}(P_{X|M} \,\|\, P_{X'|M}) \right] = 0, \end{aligned} \tag{19}$$

where the constraint $P_{X'|M} = P_{X|M}$ has been expressed in terms of zero expected KL-divergence between the channels.

Equations (18) and (19) reveal the remarkable similarity between the δ- and $\sim$-PIDs. The two PIDs are essentially optimizing over the same quantities, but in effect, interchange objective and constraint.

E. Clarifying the connection to Blackwell sufficiency, and a new informal interpretation

Using the newfound connection between the δ- and $\sim$-PIDs, we can clarify their connection to Blackwell sufficiency, and provide an informal interpretation.

First, Blackwell sufficiency can be re-understood as follows. $Y \succcurlyeq_M X$ if two requirements are met: (i) there must exist a random variable X' that is derived from Y through the stochastic transformation $P_{X'|Y}$, i.e., M—Y—X' must be a *Markov chain*; and (ii) X' must act as a "*copy*" of X w.r.t. M, in the sense that $P_{X'|M} = P_{X|M}$.[10]

When $Y \not\succcurlyeq_M X$, the δ-PID and the $\sim$-PID quantify departures from Blackwell sufficiency in two different ways (also see Figure 2): (i) the δ-PID enforces the Markov chain and measures how far we are from a copy (refer Eq. 18); (ii) the $\sim$-PID enforces the copy and measures how far we are from having a Markov chain (refer Eq. 19). This unified explanation of the δ- and $\sim$-PIDs has not been identified in the literature previously, to our knowledge.

We can also use this picture to offer a new informal interpretation. If Alice and Bob opt for X and Y respectively

[10]This is equivalent to the "simulatable" notion presented in [3, Defn. 38].

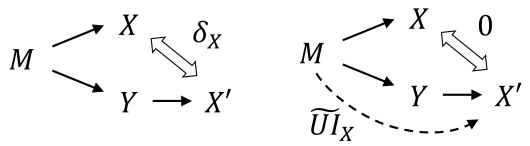

Fig. 2: A depiction of the informal interpretations of the δ- and $\sim$-PIDs, as described in Section III-E. (Left) The δ-PID enforces the Markov chain M—Y—X', and measures how far X' is from a copy of X, i.e., it measures the divergence between $P_{X|M}$ and $P_{X'|M}$. (Right) The $\sim$-PID breaks the Markov chain by allowing bits of M to leak to X' outside of Y, however, it enforces that X' is a copy of X. $\widetilde{UI}_X$ measures the minimum number of bits X' needs to borrow from M along the dashed line, so that $P_{X'|M}$ is a copy of $P_{X|M}$.

in the decision problem of Definition 5, the deficiency δ_X measures the closest that Bob can come to emulating Alice (on average, for the worst loss $\mathcal{L}$ for Bob). On the other hand, $\widetilde{UI}_X$ measures the minimum number of bits Bob needs to borrow from M in order to emulate Alice perfectly (Figure 2).

Remark. Banerjee et al. [3] also present another variant of deficiency, which they call the weighted input deficiency. They show that it induces a PID just as in Definition 1, and that this PID is identical to another early and well-known PID proposed by Harder et al. [21]. Despite the similarity between the two deficiency notions, however, the PID based on input deficiency is *not Blackwellian* [3, Ex. 28(b)]. Thus, the connection presented here cannot be extended to the PID based on input deficiency.

F. A novel generalization of the δ- and $\sim$-PIDs

The connection identified above also allows us to generalize both definitions using a single Lagrangian form:

$$\delta^\lambda(M : X \setminus Y) \coloneqq \inf_{P_{X'|MY}} \mathbb{E}_M \left[D_{KL}(P_{X|M} \,\|\, P_{X'|M}) \right] + \lambda\, I(M; X' \,|\, Y). \tag{20}$$

As $\lambda \to \infty$ in the Equation (20), we get the δ-PID, and as $\lambda \to 0$, we get the $\sim$-PID. This new δ^λ-PID has to be written in terms of a deficiency and then symmetrized as in Definition 1, since its redundancy will not be symmetric in general.

IV. Capturing the Unique Information

In Section III, we discussed operational interpretations for the quantity of unique information in two closely related PID definitions. However, both of these definitions are statistical rather than structural, i.e., they do not tell us *what* the unique information *is*. This could be relevant in a number of settings, such as for fairness in Machine Learning, as motivated in our previous work [15].

In this section, we propose a new PID definition that is able to *capture* the unique information in the form of a random variable. The quantity of unique information also has a simple operational interpretation in terms of mutual information.

Definition 6 (I-PID)**.** *Let the* information deficiency *of Y with respect to X about M be given by*

$$\delta^I(M : X \setminus Y) \coloneqq \sup_{P_{T|M} \in \mathcal{C}(\mathsf{T}|\mathsf{M})} I(T; X) - I(T; Y). \tag{21}$$

Here, T is a random variable produced through the stochastic transformation $P_{T|M}$, and satisfies the Markov chain T—M—(X, Y). Then, the redundant information may be defined as

$$RI^I(M : X; Y) = \min\big\{ I(M; X) - \delta^I(M : X \setminus Y), \; I(M; Y) - \delta^I(M : Y \setminus X) \big\}. \tag{22}$$

This definition is appealing, since it captures the basic intuition that if X has unique information about M with respect to Y, that means that X has information about some "part" of M which Y does not have access to. In practice, this could mean that X is able to access entire "dimensions" of M that Y cannot, or it could mean that X has access to some of the same dimensions of M as Y, but with lower noise, or it could be a combination of these factors. In this definition, the stochastic transformation $P_{T|M}$ plays the role of *extracting* these "parts" of M, which X has access to, but Y does not. The random variable T corresponding to the optimal $P_{T|M}$ tells us the "parts" (or subspaces) of M in which X has unique information w.r.t. Y.

The operational interpretation for the unique information of the I-PID is simply this: UI^I_X is the maximum information about M which you can extract from X, which you cannot simultaneously get from Y. That is, for any (possibly stochastic) function f that depends only on M, we will always have

$$I\big(f(M); X\big) \leq I\big(f(M); Y\big) + UI^I(M : X \setminus Y). \tag{23}$$

However, this definition also suffers from the cyan region problem described in Section III-C. This is one area where we still need to work on understanding its interpretation.

In what follows, we prove some basic properties about the I-PID, and show that it is Blackwellian for Gaussian P_{MXY}.

Theorem 4 (Non-negativity and bounds on the I-PID)**.** *The I-PID atoms can be shown to be non-negative:*

$$UI^I(M : X \setminus Y) \geq 0 \qquad RI^I(M : X; Y) \geq 0$$
$$UI^I(M : Y \setminus X) \geq 0 \qquad SI^I(M : X; Y) \geq 0$$

The I-PID also satisfies the natural bounds:

$$UI^I(M : X \setminus Y),\; RI^I(M : X; Y) \leq I(M; X),$$
$$UI^I(M : X \setminus Y),\; SI^I(M : X; Y) \leq I(M; X \,|\, Y).$$

Theorem 5 (The I-PID is Blackwellian for Gaussian P_{MXY})**.** *If P_{MXY} is jointly Gaussian, then the I-PID unique information satisfies:*

$$UI^I(M : X \setminus Y) = 0 \quad \Leftrightarrow \quad Y \succcurlyeq_M X. \tag{24}$$

Proofs of these theorems are presented in Appendix B [1]. In particular, Theorem 5 implies that prior results we have shown for Gaussian distributions [16] are also applicable to the I-PID. We conjecture that Theorem 5 can be generalized, i.e., the I-PID is Blackwellian in general, but leave an investigation of this to future work.

Also highly relevant are properties such as continuity and additivity [22]. As such, these are beyond the scope of the current paper, and we leave their examination to future work.

ACKNOWLEDGMENTS

P. Venkatesh was supported by a Shanahan Family Foundation Fellowship at the Interface of Data and Neuroscience at the Allen Institute and the University of Washington, supported in part by the Allen Institute. We wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

REFERENCES

[1] Full version of this paper with appendices. [Online]. Available: https://praveenv253.github.io/assets/doc/papers/2023--isit--full-paper.pdf
[2] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.
[3] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, "Unique informations and deficiencies," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2018, pp. 32–38, Example 28(b), as well as a corrected version of Proposition 28, may be found in the updated version on arXiv, arXiv:1807.05103v3 [cs.IT].
[4] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.
[5] G. Pica, E. Piasini, H. Safaai, C. Runyan, C. Harvey, M. Diamond, C. Kayser, T. Fellin, and S. Panzeri, "Quantifying how much sensory information in a neural code is relevant for behavior," in *Advances in Neural Information Processing Systems*, 2017, pp. 3686–3696.
[6] N. M. Timme and C. Lapish, "A tutorial for information theory in neuroscience," *eneuro*, vol. 5, no. 3, 2018.
[7] T. Scagliarini, L. Faes, D. Marinazzo, S. Stramaglia, and R. N. Mantegna, "Synergistic information transfer in the global system of financial markets," *Entropy*, vol. 22, no. 9, p. 1000, 2020.
[8] D. A. Ehrlich, A. C. Schneider, M. Wibral, V. Priesemann, and A. Makkeh, "Partial information decomposition reveals the structure of neural representations," *arXiv preprint arXiv:2209.10438*, 2022.
[9] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "An information-theoretic quantification of discrimination with exempt features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3825–3833.
[10] J. T. Lizier, N. Bertschinger, J. Jost, and M. Wibral, "Information decomposition of target effects from multi-source interactions: perspectives on previous, current and future work," p. 307, 2018.
[11] A. Kolchinsky, "A novel approach to the partial information decomposition," *Entropy*, vol. 24, no. 3, p. 403, 2022.
[12] J. Rauh, P. K. Banerjee, E. Olbrich, and J. Jost, "Unique information and secret key decompositions," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 3042–3046.
[13] E. Torgersen, *Comparison of statistical experiments*. Cambridge University Press, 1991, vol. 36.
[14] E. Mariucci, "Le cam theory on the comparison of statistical models," *arXiv preprint arXiv:1605.03301*, 2016.
[15] K. Gurushankar, P. Venkatesh, and P. Grover, "Extracting unique information through markov relations," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2022, pp. 1–6.
[16] P. Venkatesh and G. Schamberg, "Partial information decomposition via deficiency for multivariate gaussians," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 2892–2897.
[17] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided self-organization: inception*. Springer, 2014, pp. 159–190.
[18] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, pp. 265–272, 1953.
[19] L. Le Cam, "Sufficiency and approximate sufficiency," *Ann. Math. Statist.*, vol. 35, no. 4, pp. 1419–1455, 12 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177700372
[20] M. Raginsky, "Shannon meets Blackwell and Le Cam: Channels, codes, and statistical experiments," in *2011 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2011, pp. 1220–1224.
[21] M. Harder, C. Salge, and D. Polani, "Bivariate measure of redundant information," *Physical Review E*, vol. 87, no. 1, p. 012130, 2013.
[22] J. Rauh, P. K. Banerjee, E. Olbrich, G. Montúfar, and J. Jost, "Continuity and additivity properties of information decompositions," *arXiv preprint arXiv:2204.10982*, 2022.
[23] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY: Springer, 2009. [Online]. Available: http://dx.doi.org/10.1007/b13794
[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

APPENDIX A
PROOF OF THEOREM 1

Proof. Consider the difference in average risks:

$$\begin{aligned}
&\bar{\mathcal{R}}(P_{X|M}, \hat{M}_X, \mathcal{L}) - \bar{\mathcal{R}}(P_{Y|M}, \hat{M}_Y, \mathcal{L}) \qquad (25)\\
&= \mathbb{E}_M\Big[\mathbb{E}_{X|M}\big[\mathcal{L}(\hat{M}_X(X), M)\big] - \mathbb{E}_{Y|M}\big[\mathcal{L}(\hat{M}_Y(Y), M)\big]\Big]\\
&= \mathbb{E}_M\Big[\int P_{X|M}\cdot\mathcal{L}(\hat{M}_X(X), M)\,dx\\
&\qquad - \int P_{Y'|M}\circ P_{X|M}\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\\
&\qquad + \int P_{Y'|M}\circ P_{X|M}\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\\
&\qquad - \int P_{Y|M}\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\Big] \qquad (26)
\end{aligned}$$

Now, the last two terms of this expression can be bounded using the bound on $\mathcal{L}$ and the total variation distance:

$$\begin{aligned}
&\mathbb{E}_M\Big[\int P_{Y'|M}\circ P_{X|M}\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\\
&\qquad - \int P_{Y|M}\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\Big] \qquad (27)\\
&= \mathbb{E}_M\int\big(P_{Y'|M}\circ P_{X|M} - P_{Y|M}\big)\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\\
&\overset{(a)}{\leq} \|\mathcal{L}\|_\infty\cdot\mathbb{E}_M\left\|P_{Y'|M}\circ P_{X|M} - P_{Y|M}\right\|_{TV} \qquad (28)\\
&\overset{(b)}{\leq} \|\mathcal{L}\|_\infty\cdot\frac{1}{\sqrt{2}}\,\mathbb{E}_M\sqrt{D_{KL}\big(P_{Y|M}\big\|P_{Y'|M}\circ P_{X|M}\big)} \qquad (29)\\
&\overset{(c)}{\leq} \|\mathcal{L}\|_\infty\cdot\sqrt{\frac{1}{2}\,\mathbb{E}_M\,D_{KL}\big(P_{Y|M}\big\|P_{Y'|M}\circ P_{X|M}\big)} \qquad (30)\\
&\overset{(d)}{=} \|\mathcal{L}\|_\infty\cdot g\big(\delta(M:Y\setminus X)\big), \qquad (31)
\end{aligned}$$

where in (a) we have used the bound on $\mathcal{L}$ and the definition of the total variation norm, in (b) we have used Pinsker's inequality [23, Lemma 2.5], in (c) we have used Jensen's inequality [24, Thm. 2.6.2], and in (d), we have set $g(z) := \sqrt{z/2}$.

It only remains to be shown that the first two terms of the expression in Equation (26) can be upper bounded by zero. Examining the first two terms, for any $\hat{M}_Y(y)$, we can derive a stochastic action rule, $\hat{M}_X(x)$ that will attain the same risk: we can first draw $\tilde{y}\sim P_{Y'|X}$ and then select the action $\hat{M}_Y(\tilde{y})$. Thus,

$$\begin{aligned}
&\mathbb{E}_M\Big[\int P_{X|M}\cdot\mathcal{L}(\hat{M}_X(X), M)\,dx \qquad (32)\\
&\qquad - \int P_{Y'|M}\circ P_{X|M}\cdot\mathcal{L}(\hat{M}_Y(Y), M)\,dy\Big] \leq 0,
\end{aligned}$$

which completes the proof. □

Appendix B
Proofs of Theorems 4 and 5

Proof of Theorem 4. First, observe that

$$\delta^I(M : X \setminus Y) \coloneqq \sup_T I(T;X) - I(T;Y) \tag{33}$$
$$\geq I(0;X) - I(0;Y) = 0. \tag{34}$$

Furthermore,

$$I(T;X) - I(T;Y) \leq I(T;X) \leq I(M;X), \tag{35}$$

where the last inequality follows by the data processing inequality and the Markov chain T—M—(X,Y). Thus,

$$\delta^I(M : X \setminus Y) \leq I(M;X) \tag{36}$$
$$0 \leq I(M;X) - \delta^I(M : X \setminus Y) \leq I(M;X) \tag{37}$$
$$0 \leq I(M;Y) - \delta^I(M : Y \setminus X) \leq I(M;Y) \tag{38}$$

This implies

$$0 \leq RI^I(M : X;Y) \leq \min\{I(M;X), I(M;Y)\} \tag{39}$$
$$0 \leq UI^I(M : X \setminus Y) \leq I(M;X) \tag{40}$$
$$0 \leq UI^I(M : Y \setminus X) \leq I(M;Y) \tag{41}$$

Furthermore,

$$I(T;X) - I(T;Y) = I(T;(X,Y)) - I(T;Y \,|\, X) \tag{42}$$
$$- I(T;(X,Y)) + I(T;X \,|\, Y)$$
$$= I(T;X \,|\, Y) - I(T;Y \,|\, X) \tag{43}$$
$$\leq I(T;X \,|\, Y) \leq I(M;X \,|\, Y), \tag{44}$$

where in the very last inequality follows from the fact that $T \perp\!\!\!\perp (X,Y) \,|\, M$ and the data processing inequality [24, Ch. 2]. This may not be obvious, but it follows the same proof as the data processing inequality:

$$I\big(T,M;X \,|\, Y\big) = I\big(T;X \,|\, Y\big) + I\big(M;X \,|\, Y,T\big) \tag{45}$$
$$= I\big(M;X \,|\, Y\big) + I\big(T;X \,|\, Y,M\big) \tag{46}$$

From this it follows that

$$I\big(T;X \,|\, Y\big) + I\big(M;X \,|\, Y,T\big) \overset{(a)}{=} I\big(M;X \,|\, Y\big) \tag{47}$$
$$I\big(T;X \,|\, Y\big) \overset{(b)}{\leq} I\big(M;X \,|\, Y\big), \tag{48}$$

where (a) follows from the fact that $I\big(T;X \,|\, Y,M\big) = 0$ since $T \perp\!\!\!\perp (X,Y) \,|\, M$, while (b) uses $I\big(M;X \,|\, Y,T\big) \geq 0$. This justifies Equation (44), which implies

$$\delta^I(M : X \setminus Y) \leq I(M;X \,|\, Y) \tag{49}$$
$$\delta^I(M : Y \setminus X) \leq I(M;Y \,|\, X) \tag{50}$$

If $UI^I(M : X \setminus Y) = \delta^I(M : X \setminus Y)$, then $SI^I(M : X;Y) = I(M;X \,|\, Y) - \delta^I(M : X \setminus Y) \geq 0$, and $SI \leq I(M;X \,|\, Y)$. This shows that all terms in the I-PID are non-negative and bounded. □

Proof of Theorem 5. We need to show that when P_{MXY} is jointly Gaussian,

$$UI^I_X = 0 \quad \Leftrightarrow \quad Y \succcurlyeq_M X. \tag{51}$$

($\Leftarrow$) Observe that the I-PID satisfies Assumption $(*)$ from Bertschinger et al. [4], i.e., UI_X is a function only of P_M, $P_{X|M}$ and $P_{Y|M}$. Thus, by [4, Lemma 3], $UI^I_X \leq \widetilde{UI}_X$. Since the $\sim$-PID is Blackwellian, $Y \succcurlyeq_M X \Leftrightarrow \widetilde{UI}_X = 0 \Rightarrow UI^I_X = 0$.

This part of the proof holds irrespective of the distribution of P_{MXY}.

($\Rightarrow$) Now, suppose P_{MXY} is Gaussian. Then it suffices to show that whenever $Y \not\succcurlyeq_M X$, $\exists\, P_{T|M}$ such that $I(T;X) - I(T;Y) > 0$, to ensure that $UI^I_X > 0$.

Following the notation of [16], let Σ_{MXY} be represent the joint covariance matrix (which fully specifies information measures on the joint distribution), let $\Sigma_{X|M}$ represent the conditional covariance matrix of X given M and let $\Sigma_{X,Y}$ represent the cross-covariance of X and Y. Let $\Lambda_X \coloneqq \Sigma^{\mathsf{T}}_{X,M}\Sigma^{-1}_{X|M}\Sigma_{X,M}$ and $\Lambda_Y \coloneqq \Sigma^{\mathsf{T}}_{Y,M}\Sigma^{-1}_{Y|M}\Sigma_{Y,M}$. Then, [16, Theorem 2], states

$$Y \succcurlyeq_M X \quad \Leftrightarrow \quad \Lambda_Y \succcurlyeq \Lambda_X, \tag{52}$$

where for positive semidefinite matrices A and B, $A \succcurlyeq B$ denotes that $A - B$ is positive semidefinite.

Consider $P_{T|M}$ to be a normal distribution, given by $\mathcal{N}(H_T M, \Sigma_{T|M})$. Further, we can assume without loss of generality that $\Sigma_M = I$. Then, $\Sigma_{T,X} = H_T\Sigma_M\Sigma_{M,X} = H_T\Sigma^{\mathsf{T}}_{X,M}$. The mutual information between T and X is given by:

$$I(T;X)$$
$$= \frac{1}{2}\log\det(I + \Sigma_T^{-1}H_T\Sigma^{\mathsf{T}}_{X,M}\Sigma^{-1}_{X|M}\Sigma_{X,M}H_T^{\mathsf{T}}) \tag{53}$$
$$= \frac{1}{2}\log\det(I + \Sigma_T^{-1/2}H_T\Sigma^{\mathsf{T}}_{X,M}\Sigma^{-1}_{X|M}\Sigma_{X,M}H_T^{\mathsf{T}}\Sigma_T^{-1/2}) \tag{54}$$
$$= \frac{1}{2}\log\det(I + \Sigma_T^{-1/2}H_T\Lambda_X H_T^{\mathsf{T}}\Sigma_T^{-1/2}) \tag{55}$$

Then,

$$\delta(M : X \setminus Y) = \frac{1}{2}\log\det(I + \Sigma_T^{-\frac{1}{2}}H_T\Lambda_X H_T^{\mathsf{T}}\Sigma_T^{-\frac{1}{2}}) \tag{56}$$
$$- \frac{1}{2}\log\det(I + \Sigma_T^{-\frac{1}{2}}H_T\Lambda_Y H_T^{\mathsf{T}}\Sigma_T^{-\frac{1}{2}})$$

If $Y \not\succcurlyeq_M X$, then $\Lambda_Y \not\succcurlyeq \Lambda_X$, i.e., $\exists\, c \in \mathbb{R}$ s.t.

$$c^{\mathsf{T}}\Lambda_X c > c^{\mathsf{T}}\Lambda_Y c. \tag{57}$$

Letting $\Sigma_T = I$ and $H_T = c$, we have that

$$1 + c^{\mathsf{T}}\Lambda_X c > 1 + c^{\mathsf{T}}\Lambda_Y c \tag{58}$$
$$\det(1 + c^{\mathsf{T}}\Lambda_X c) > \det(1 + c^{\mathsf{T}}\Lambda_Y c) \tag{59}$$
$$\frac{1}{2}\log\det(1 + c^{\mathsf{T}}\Lambda_X c) > \frac{1}{2}\log\det(1 + c^{\mathsf{T}}\Lambda_Y c) \tag{60}$$

This implies

$$\frac{1}{2}\log\det(1 + c^{\mathsf{T}}\Lambda_X c) - \frac{1}{2}\log\det(1 + c^{\mathsf{T}}\Lambda_Y c) > 0 \tag{61}$$
$$\Rightarrow \qquad \delta(M : X \setminus Y) > 0 \tag{62}$$

Recognizing that $UI^\delta(M : X \setminus Y) \geq \delta(M : X \setminus Y)$ (see Equation (15)), this completes the proof. □